Human Action Recognition in Smart Classroom

Haibing Ren Guangyou Xu
(Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, P.R.China)
E-mail: renhaibing@263.net xgy-dcs@mail.tsinghua.edu.cn

Abstract

This paper presents a new system for teachers' natural complex action recognition in smart classroom in order to realize intelligent cameraman and virtual mouse. First, the system proposes a hybrid human model and employs 2-order B-spline function to detect the two shoulder joints in the silhouette image to obtain the basic motion features including the elbow angles, motion parameters of the face and two hands. Then, Primitive-based Coupled Hidden Markov Model (PCHMM) is presented for natural context-dependent action recognition. Last, some comparison experiments show PCHMM is better than the traditional HMM and coupled HMM.

Keywords: action recognition, primitive features, Coupled Hidden Markov Model

1. Introduction

In recent years, the research on intelligent environment connection with ubiquitous computing and pervasive computing has attracted more and more attention, such as EasyLiving(Microsoft), Intelligent Rooms(MIT AI Lab) and KidsRoom(MIT Media Lab). There arises the need for the computers to detect the subject in environment, and further to recognize him, understand his intention as well as behaviors and adapt to his habits, which is often called "Looking at People". As a result, a number of research areas should be involved, including face detection, expression recognition, gesture recognition, human tracking, pose estimation, body language understanding, etc.

The Smart classroom is a project of intelligent environments for tele-education. Almost all the present tele-education systems require the teachers to sit down in front of the video camera, however the teacher's experience is of much difference from teaching in an ordinary classroom. The smart classroom is a virtual classroom, where there is a 'blackboard' (media board), students (student icons and student video) and others, like a real one. And by the technology of human computer interaction and augment reality, the teacher feel very comfortable and could also use speech, gesture, body language and handwriting to improve efficiency as well.

This paper focuses on the teacher's upper-limb action recognition to understand the teacher's intention for intelligent cameraman and virtual mouse system. The framework of the recognition system is as following:

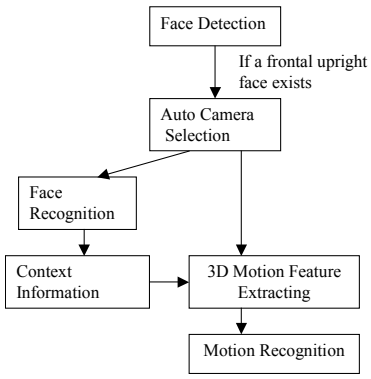


Figure 1. System framework

In the smart classroom, there are a lot of cameras around the teacher. By the face detection module [3], the two neighboring cameras that face the teacher are auto-selected. Therefore, only the frontal actions are considered in this paper. By face recognition module[6], the teacher's identity is recognized and his personal information could be retrieved in the database, which is very important to get more 3D motion feature in section 2 and adapt his habit in further application.

1.1 Related Work

In most previous research on action recognition, the actions are confined to a predefined command set, which requires that the subjects are well trained and the motions are uniform. The features and recognition algorithms are totally data-driven without any

information of high-level feature and context information.

For example, in [7], moment-based features are extracted from multiple views of motion energy images(MEI) and motion history images(MHI), and template matching algorithm are employed to recognize the aerobics exercises and the well performed moves in the KidsRoom. And in [2], with the coordinates of the two hands, Hidden Markov Models are used to recognize American sign languages. In [9], with 3D trajectories of the two hands, coupled Hidden Markov Models are presented to recognize 3 kinds of T'ai Chi Ch'uan.

However, the teacher's natural action recognition is much more difficult than the above. The motion is more natural, complex and dependent on the context and scenario. No subject would be trained and everyone has his own habit.

1.2 Our approach

In psychological research on action recognition, it was found that motion models in subjects mind are not the motion parameters such as the parameters of position and motion speed, but only some characteristic features, like the relative position of hands and face, the relation between the moving hand and the scenario, etc.

Based on this principle, this paper presents a new system to recognize the teacher's natural complex action in the smart classroom. And much of it is general and can be used in other areas.

First, this paper gives the hybrid human model to obtain the basic motion features including motion feature of the face, two hands and the two elbow angles. Then, a recognition algorithm, named Primitive-based Coupled Hidden Markov Model (PCHMM), is presented to recognize the subject's natural complex action.

As a totally data-driven algorithm, the traditional CHMM is not suit to natural complex action recognition because the feature dimension is too large and the within-class scatter is too much. And unfortunately, it is impossible to get enough training samples containing everyone's every type of action. Unlike traditional CHMM, PCHMM is an approach with high-level features and context information, which need less training samples and could diminish the within-class scatter greatly.

This paper is arranged as follows: section two describes the hybrid human model and the basic motion feature estimation; and section three gives the PCHMM for upper-limb action recognition; section four is the experimental results and last is the conclusion.

2. Hybrid Human Model

In many action recognition systems, only the trajectories of two hands (or only 1 hand) are extracted as motion features, hence much ambiguity in recognition.

In contrast, this paper presents the hybrid human model as the Figure 2 to get more explicit 3D motion features. This model includes the face, the trunk, two arms and two hands, which are indispensable parts to any person and are easily detected, and with which more information can be given to reduce the complexity and improve the robustness greatly.

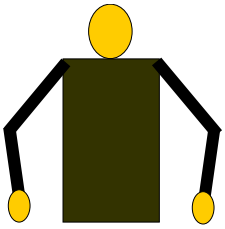


Figure 2. Hybrid human model

This model is a hybrid from 2D model and 3D model. The face, two hands are 2D ellipses in 3D space and the trunk is a 2D rectangle in 3D space. On the other hand, the upper and lower arms are 3D cylinders.

2.1 Context

In this paper, the context is a generic term, including the teacher's personal information indexed by the face recognition result. Though only a little is useful in this paper, the teacher's model is very important for further applications such as human 3D modeling, more precise motion estimation, habit self-adaptation, etc.

The context also includes some scenario information, such as the position of some objects on the desk. And with the result of previous action understanding (taking objects from the desk, putting back objects, etc), the variant b_{object} in the context represents the object in the subject's hand.

2.2 Shoulder joints detection

With the teacher's silhouette image obtained by background subtraction, this paper gives an effective approach to detect the two shoulder joints. First, the trunk area is segmented in the silhouette image as following:

$$Area_{Trunk} = C(Area_{Silhouette} - Area_{Face}) - Area_{arm} \quad (1)$$

where the $Area_{Silhouette}$ is the silhouette area as Figure 3(a), $Area_{Face}$ is the face area , $Area_{arm}$ is the two arm area(including two hands) and C is an operator to get the maximum connective area. The histogram of C($Area_{Silhouette}$ - $Area_{Face}$) on the X coordinate as Figure 3(c) can be looked as 5 lines as Figure 3(d). The areas covered by l_1 and l_5 are considered as the two arms area $Area_{arm}$. By deleting the $Area_{arm}$, the left is $Area_{Trunk}$ as Figure 3(e). The parameters of l_i ($i = 1,...,5$) is estimated to satisfy the following constraint:

$$\min \int_{s=0}^{1} (|H_x(s) - F_x(s)| + |H_y(s) - F_y(s)|) \quad (2)$$

where s is arc length normalized to $[0,1]$, $H(s) = (H_x(s), H_y(s))$ is the histogram function, $F(s) = (F_x(s), F_y(s))$ is the function composed of the 5 lines. The equation (2) is a global optimization problem with hyper parameters. In this paper, 2-order B-spline with 5 segments is employed to get the solution and the result is as the Figure 3(d), which is very precise, robust and only costs 2.56ms.

Last, two special corner operators are used to detected the two shoulder joints in the ROI (region of interest) which is determined by the position and the size of the $Area_{Face}$. The detection of the shoulder joint is based the global information. Though its 3D coordinates aren't very precise, they are very robust.

When a hand is in front of the trunk, there may be some ambiguity between l_1 and l_2 (or l_4 , l_5). But in this case, l_1 and l_2 are very small in the histogram and no matter how much the error of $Area_{arm}$ is, the deviation of $Area_{Trunk}$ will be very little.

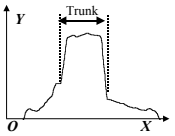
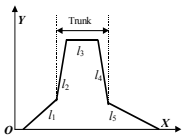
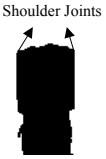


Figure 3.(a) $Area_{Silhouette}$ **(b)** **C(** $Area_{Silhouette}$ - $Area_{Face}$**)** **(c) *H*(*s*)**



(d) *F*(*x*)



(e) $Area_{Trunk}$

2.3 Basic motion features

With the detection of the face, two hands and two shoulder joints in stereo images, their 3D coordinates could be obtained. Then elbow joint θ ($0 \le \theta \le 90^\circ$) are calculated with the integration of shoulder joint 3D coordinate and the length of the upper, lower arms. In this paper, the elbow angle is more important than elbow joint 3D coordinate. The reason is:

(1) Because human 3D model building is time-consuming and unstable, it is nearly impossible to estimate the elbow joint 3D coordinate precisely and robustly without any markers.
(2) For action recognition, it isn't necessary to get the elbow joint 3D coordinates because the position of elbow joint is meaningless in most human motion.
(3) The elbow angle is significant to represent the arm state.

Therefore, the basic motion feature for each hand is obtained as following:

$$(P_{hand}, V_{hand}, A_{hand}, P_{face}, V_{face}, \theta_{elbow}, b_{object}) \quad (3)$$

where P_{hand} , V_{hand} , A_{hand} is the hand 3D position, 3D velocity and 3D acceleration respectively. And P_{face} , V_{face} is Face 3D position and 3D velocity, θ_{elbow} is the elbow angle and b_{object} is a variant in context. From equation (3), the following could be got:

$$V_{hand} = \partial P_{hand} \quad A_{hand} = \partial V_{hand} \quad (4)$$

$$V_{face} = \partial P_{face} \quad (5)$$

Due to different teachers and different actions, there is much within-class scatter in the 17-dimension basic motion features. And it could not be used directly for action recognition.

3. Primitive-based Coupled Hidden Markov Model

This paper introduces the primitive features to the traditional coupled-HMM and calls it Primitive-based Coupled Hidden Markov Model.

3.1 Primitive features

The states in action recognition often have definite meaning and much clear segmentation. This is because each state has some unambiguous features, called primitive features or primitives in this paper. Each primitive feature λ is represented by a Gaussian Mixture Model (*GMM*) and the distribution density is as the following:

$$P(p \mid \lambda) = \sum_{i=1}^{G_n} G_i(p) P(i) \quad (6)$$

where p is a primitive variant, $G_i(p)$ (i=1,…, G_n) is a Gaussian model with mean ξ_i and covariance matrix σ_i, which are hyperparameters of the distribution, G_n is the Gaussian model number, $p(i)$ is the weight function for each Gaussian model. The parameters of $G_i(p)$ and $p(i)$ could be estimated by Expectation Maximum (EM) algorithm.

3.2 Representation of the states

The states in this paper is strictly defined by some primitive features and corresponding weight. These primitive features of state S are supposed to be independent of each other. The observation densities function of S is as following:

$$P(O \mid S) = P(P \mid S) = \sum_{i=1}^{P_n} P(p_i \mid \lambda_i) W(i) \quad (7)$$

where O is the observation(basic motion feature in this paper), P_n is the primitive feature number of S, λ_i is the ith primitive feature (i=1,…, P_n), $W(i)$ is the corresponding weight for λ_i, $P = \{p_1, \cdots, p_{p_n}\}$ is the primitive set of S. And the relation of p_i and O can be described as the equation (7):

$$p_i = f_i(O, S, Context) \quad (8)$$

where *Context* is the context information and f_i is the function to extract the p_i from O, S and *Context*. In this paper, $W(i)$ is estimated by the following:

$$\begin{cases} \arg\max_{W(1),W(2),\ldots,W(T_n)} \left[\sum_{i=1}^{T_n} P(O_i \mid S) \right] \\ \sum_{i=1}^{M} W(i) = 1 \end{cases} \quad (9)$$

where *W(i)*, S are the same as equation(6), O_i is the ith training sample for state S and T_n is the number of the training samples. Here, $W(i)$ ($i = 1,\ldots,T_n$) is estimated by maximum likelihood.

3.3 Primitive-based Coupled Hidden Markov Model

For each hand, the basic motion feature from time 1 to T is considered to be 1-order Markov chain. And suppose that the relation between the two hands satisfy PCHMM, as the Figure 4.

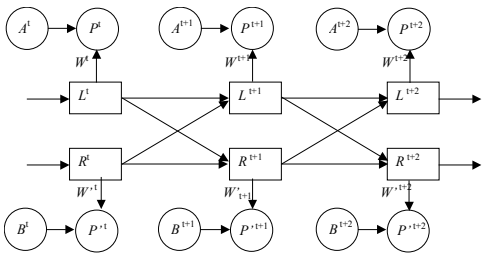


Figure 4. PCHMM structure.

where the superscript $t, t+1, t+2$ in the structure mean at time t, $t+1$, $t+2$ respectively, L^t is the left hand state at time t, A^t is the basic motion feature of the left hand at time t, P^t is the primitive feature of left hand motion, W^t is the weight vector for L^t, R^t is the right hand state at time t, B^t is the basic motion feature sequence of the right hand at time t, P'^t is the primitive feature of the right hand motion, W'^t is the weight vector for R^t.

$L_1^T = \{L^t, t = 1,\ldots,T\}$ is the left hand state sequence, $A_1^T = \{A^t, t = 1,\ldots,T\}$ is the basic motion feature sequence of the left hand. $R_1^T = \{R^t, t = 1,\ldots,T\}$ is the right hand state sequence, $B_1^T = \{B^t, t = 1,\ldots,T\}$ is the basic motion feature sequence of the right hand. The likelihood function of PCHMM with the basic motion feature A_1^T and B_1^T is:

$$\begin{aligned} P(A_1^T, B_1^T \mid \Theta) &= P(L_1^T) * P(R_1^T) * P(A^1 \mid L^1) \\ &\quad P(A^t \mid L^t) * P(B^t \mid R^t) \\ &* P(B^1 \mid R^1) * \prod_{t=2}^{T} * P(A^t \mid L^{t-1}, R^{t-1}) * \\ &\quad P(B^t \mid L^{t-1}, R^{t-1}) \end{aligned} \quad (10)$$

where the PCHMM parameter set Θ contains the prior probabilities $P(L_1^T)$ and $P(R_1^T)$ for the two Markov chains, the observation densities function $P(A^t \mid L^t)$ and $P(B^t \mid R^t)$, the transition probabilities $P(L^t \mid L^{t-1}, R^{t-1})$ and $P(R^t \mid L^{t-1}, R^{t-1})$. The prior probabilities $P(L_1^T)$ and $P(R_1^T)$ are supposed to be equal for each model. With forward-backward Viterbi algorithm, the parameters set Θ of the model can be estimated as following:

$$\arg\max_{\Theta} \sum_{i=1}^{S_n} P(A_{i1}^T, B_{i1}^T \mid \Theta) \qquad (11)$$

where S_n is the sequence number for training, (A_{i1}^T, B_{i1}^T) is the *i*th training sequence.

4. Experimental Results

For the teacher in the smart classroom, there are totally 7 kinds of natural actions to be recognized:

- Taking objects from the desk
- Putting back objects
- Pointing to the students
- Pointing to the blackboard (virtual mouse)
- Communication with the students
- Explaining objects
- Drinking water

For each action, there are 50 samples. Comparison experiments are done among HMM, traditional CHMM and PCHMM and the result is as following:

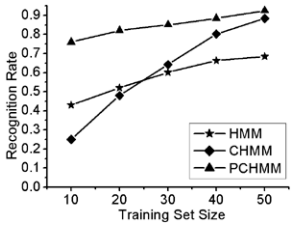


Figure 5. The comparison among HMM, traditional CHMM and PCHMM

where the *x* coordinate is the size of the training data set and all the testing sets are the whole data set. It shows PCHMM is the best among the three algorithms, especially with less training data.

Another comparison experiments on the elbow angles are carried out. And the result is as following:

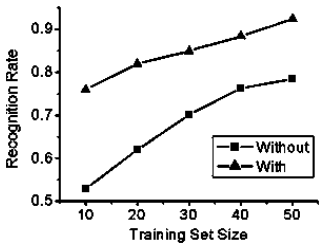


Figure 6. The comparison on the elbow angles

where '**With**' means motion features include the elbow angles and '**Without**' means the reverse. The Figure 6 shows the basic motion features with the elbow angles perform much better than the one without them. Though the elbow angles are not very precise, they are very important features to represent the arm state. At the same time, it shows the face and two hands motion features aren't sufficient for complex action recognition.

5. Conclusion

This paper presents a framework for the teacher's complex action recognition in the smart classroom. With the Hybrid Human Model, basic motion feature are extracted which includes the two elbow angles and the motion features of the head and two hands. Primitive-based Coupled-HMM are used for recognition. And the encouraging experiment result show the PCHMM is very robust and can obtain better result especially in the case of only less training samples.

Finally, the prototype of the smart classroom for tele-education is as following:

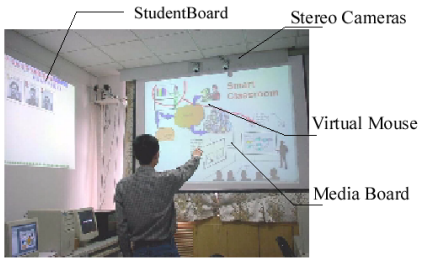


Figure 7.The prototype of smart classroom.

In this prototype, the framework of natural complex action recognition performs the smart cameraman and virtual mouse very robust and accurately.

5. References

[1] G. Johansson. Visual perception of biological motion and a model for its analysis. Perception and Psychophysics, 14(2):201-211,1973.

[2] Starner T, Weaver J, Pentland. A real-time American sign language recognition using desk and wearable computer based video. IEEE Transaction on Pattern Analysis and Machine Intelligence [J], 1998, 20(12): 1371-1375.

[3] Haizhou Ai, Luhong Liang, Guangyou Xu, Face detection in template matching constrained subspace. In Proceedings, Edited by H.R. Arabnia, International Conference on Artificial Intelligence 2001 (IC-AI 2001), Vol.II, pp.603-608, Las Vegas, Nevada, USA, June 25-28, 2001

[4] Rezek, L., Sykacek, P., Roberts, S.J. Coupled hidden Markov models for biosignal interaction modeling. Advances in Medical Signal and Information Processing, 2000. First International Conference on (IEE Conf. Publ. No.476), 2000 Page(s): 54 –59

[5] Matthew Brand, Nuria Oliver and Alex Pentland. Coupled hidden markov models for complex action recognition. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1997: 994-999

[6] Z.Y.Peng, W.Hong, L.H.Liang, G.Y.Xu and H.J.Zhang. Detecting facial features on image sequences using cross-verification mechanism. Proceeding of the Second IEEE Pacific-Rim Conference on Multimedia, 2001, pp. 1060-1065

[7] James W. Davis, Aaron F. Bobici. The Representation and recognition of action using temporal templates. Proceeding of the International Conference on Computer Vision and Pattern Recognition, 1997, 928~934

[8] Christopher R. Wren, Brian P. Clarkson, Alex Pentland. Understanding purposeful human motion. Proceeding of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000,pp: 378-383.

[9] Matthew Brand, Nuria Oliver, Alex Pentland. Coupled hidden markov models for complex action recognition. Proceeding of IEEE Society Conference on Computer Vision and Pattern Recognition, 1997, pp: 994-999